Bài-tập-Map Reduce

Bigdata (Trường Đại học Công nghệ thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh)

Câu 1. Cho dữ liệu là những tập tin nhật ký duyệt web có cấu trúc như sau: <đường dẫn>, <thời_gian_tính_bằng_phút>

Ví dụ:
https://facebook.com, 86
https://youtube.com, 33
https://tuoitre.vn, 25
https://tinhte.vn, 22
a. Thiết kế các hàm trên mô hình MapReduce thống kê thời lượng duyệt web của người dùng theo từng trang.
b. Sử dụng 02 kỹ thuật là tổng hợp cục bộ (Combiner) trong lớp Mapper và duy trì trạng thái biến nhớ trên toàn bộ các tác vụ map, để tối ưu hóa lớp Mapper

Bài làm:

a. Class MAPPER

Method Map (id i , line ln):

```
arr <- ln.split(“,”)
min <- (int) arr[1]
Emit(arr[0], min);
```

Class REDUCE

Method Reduce(link l, minutes[m1, m2,…]):

```
Sum <- 0
for all minute m ∈ minutes[m1, m2, …] do
    sum += m;
Emit(link l, minute sum);
```

b.

* Tổng hợp cục bộ

Class MAPPER

Method Map(id I, line ln)

```
H <- new ASSOCIATIVEARRAY()
arr <- ln.split(“,”)
```

```
        l <- arr[0]
        min <- StringtoInt(arr[1])
        H{l} <- H{l} + min
        for all link l ∈ H do
                Emit (link l, minute H{l})

Class REDUCE
Method Reduce(link l, minutes[m1, m2,…]):
        Sum <- 0
        for all minute m ∈ minutes[m1, m2, …] do
                sum += m;
        Emit(link l, minute sum);
```

* Duy trì trạng thái biến nhớ

```
Class MAPPER
Method INITIALIZE:
        H <- new ASSOCIATIVEARRAY()

Method Map( id I, line ln)
        arr <- ln.split(“,”)
        l <- arr[0]
        min <- StringtoInt(arr[1])
        H{l} <- H{l} + min

Method CLOSE
        for all link l ∈ H do
                Emit (link l, minute H{l})
```

Class REDUCE

Method Reduce(link l, minutes[m1, m2,…]):

    Sum <- 0

    for all minute m ∈ minutes[m1, m2, …] do

        sum += m;

    Emit(link l, minute sum);

Câu 2. Cho dữ liệu bán hàng của một siêu thị dưới dạng tập tin có cấu trúc như sau: <mã_khách_hàng>, <mã_sản_phẩm>, <đơn_giá>, <số_lượng>

Ví dụ:

KH11321, SP110, 136000, 9

KH11321, SP092, 18500, 24

KH09231, SP003, 20300, 7

KH01287, SP206, 94300, 10

a. Thiết kế các hàm trên mô hình MapReduce tính doanh số bán hàng của siêu thị đó.

b. Sử dụng 02 kỹ thuật là tổng hợp cục bộ (Combiner) trong lớp Mapper và duy trì trạng thái biến nhớ trên toàn bộ các tác vụ map, để tối ưu hóa lớp Mapper.

Bài làm:

a. Class MAPPER

Method Map (id i , line ln):

    arr <- ln.split(“,”)

    price <- (int) arr[2]

    quantity <- (int) arr[3]

    revenue <- price*quantity

    Emit(arr[0], revenue);

```
Class REDUCE
Method Reduce(CustomerID c, revenues [r1, r2,…]):
        sum <- 0
        for all revenue r ∈ revenues[m1, m2, …] do
                sum += r;
        Emit(CustomerID c, revenue sum);
```

b.

* Tổng hợp cục bộ

```
Class MAPPER
Method Map( id I, line ln)
        H <- new ASSOCIATIVEARRAY()
        arr <- ln.split(“,”)
        price <- (int) arr[2]
        quantity <- (int) arr[3]
        revenue <- price*quantity
        H{l} <- H{l} + revenue
        for all link l ∈ H do
                Emit (CustomerID c, revenue H{l})

Class REDUCE
Method Reduce(CustomerID c, revenues [r1, r2,…]):
        sum <- 0
        for all revenue r ∈ revenues[m1, m2, …] do
                sum += r;
        Emit(CustomerID c, revenue sum);
```

* Duy trì trạng thái biến nhớ

```
Class MAPPER
Method INITIALIZE:
        H <- new ASSOCIATIVEARRAY()

Method Map( id I, line ln)
        arr <- ln.split(“,”)
        price <- (int) arr[2]
        quantity <- (int) arr[3]
        revenue <- price*quantity
        H{l} <- H{l} + revenue

Method CLOSE
        for all link l ∈ H do
                Emit (CustomerID c, revenue H{l})

Class REDUCE
Method Reduce(CustomerID c, revenues [r1, r2,…]):
        sum <- 0
        for all revenue r ∈ revenues[m1, m2, …] do
                sum += r;
        Emit(CustomerID c, revenue sum);
```

Câu 3. Cho dữ liệu đo đạc của các cảm biến không khí dưới dạng tập tin có cấu trúc như sau: <mã_cảm_biến>, <chỉ_số>, <độ_lớn>

Ví dụ:

1000120, SO2, 3.42

1000120, CO, 2602

1000120, PM10, 88.33

1000124, CO, 1358

Thiết kế các hàm trên mô hình MapReduce tính trung bình nồng độ khí CO của các cảm biến đo đạc được.

Bài làm:

```
Class MAPPER
Method Map (id i , line ln):
    arr <- ln.split(“,”)
    value <- (float) arr[2]
    if(arr[1]=”CO”)
        Emit(arr[0], (value, 1));

Class REDUCE
Method Reduce(SensorId s, values [v1, v2,…]):
    sum <- 0
    count <- 0
    average <- 0
    for all value v, number n ∈ values[v1, v2, …] do
        sum += v
        count +=n
    if(count > 0)
        average <- sum/count
    Emit(SensorId s, average);
```

Câu 4. Cho một dãy số chứa trong tập tin như sau:

18229

2902

45

3903

9539

33893

5

350

Thiết kế các hàm trên mô hình MapReduce tính trung bình nhân của dãy số trên.

Bài làm:

Class MAPPER

Method Map (id i , line ln):

```
arr <- ln.split(“,”)
Emit(id i, (arr[0], 1));
```

Class REDUCE

Method Reduce(id i, values [v1, v2,…]):

```
multiplication <- 1
count <-0
average <- 0
for all value v, count c ∈ values[v1, v2, …] do
    multiplication *= v
    count += c
if (count > 0)
    average = multiplication/count
Emit(id i, average);
```

Câu 5. Cho dữ liệu ghi nhận lượng điện năng tiêu thụ của từng hộ gia đình trong một tháng dưới dạng tập tin có cấu trúc như sau: <mã_hộ_gia_đình>, <số_điện_tiêu_thụ>

Ví dụ:

1000120, 325

3020180, 121

0277529, 78

0983910, 502

Thiết kế các hàm trên mô hình MapReduce tính tổng số tiền thu được trong tháng, biết các mức giá điện được tính theo công thức sau:

- Bậc 1: Cho kWh từ 0 – 100 là 1.734 đ
- Bậc 2: Cho kWh từ 101 – 200 là 2.014 đ
- Bậc 3: Cho kWh từ 201 – 300 là 2.536 đ
- Bậc 4: Cho kWh từ 301 trở lên là 2.834 đ

Bài làm:

Class MAPPER

Method Map (id i , line ln):

```
arr <- ln.split(“,”)
value <- (int) arr[1]
if(value <= 100)
        Emit(arr[0], value*1.734);
else if(value >= 101 and value <= 200)
        Emit(arr[0], value*2.014);
else if(value >= 201 and value <= 300)
        Emit(arr[0], value*2.536);
else if(value >= 301)
        Emit(arr[0], value*2.834);
```

Class REDUCE

Method Reduce(FamilyId f, values [v1, v2,…]):

sum <- 0

for all value v ∈ values[v1, v2, …] do

sum += v

Emit(FamilyId f, average);

Câu 6. Từ kết quả của bài tập 1, thiết kế các hàm trên mô hình MapReduce để tìm ra website người dùng dành nhiều/ít thời gian xem nhất.

Bài làm:

Câu 7. Cho dữ liệu thông tin nhân viên dưới dạng tập tin có cấu trúc như sau: <họ>, <tên>, <chức_vụ>, <phòng_ban>, <lương>

Ví dụ:

Nguyễn Hoàng, Anh, Quản lý, Hành chính, 18650000

Trương Thị Anh, Thư, Nhân viên, Nghiên cứu, 12600000

Võ Nguyễn Thùy, Trang, Nhân viên, Nhân sự, 11680000

Trần Thanh, Nhân, Quản lý, Marketing, 20070000

Lý Thị Ngọc, Diễm, Nhân viên, Marketing, 13970000

Thiết kế các hàm trên mô hình MapReduce tính số lượng người, người có lương cao nhất/thấp nhất của từng phòng ban.

Bài làm:

Câu 8. Cho dữ liệu mô tả về hình đa giác (tam giác, tứ giác, ngũ giác, lục giác …) gồm các cạnh,

được lưu trữ trong tập tin như sau:

A--B; B--C; C--A

G--H; H--I; I--K; K--G

X--Y; Y--Z; Z--X

M--N; N--O; O--P; P--Q; Q--M

Cho mã giả các lớp Mapper và Reducer của chương trình thống kê số cạnh của đa giác có trong tập tin, như sau:

```
class MAPPER
    method MAP(polygonId i, polygon p)
        for all edge e ∈ polygon p do
            EMIT(edge e, count 1)
class REDUCER
    method REDUCE(edge e, counts[c1, c2,…])
        sum ← 0
        for all count c ∈ counts[c1, c2,…] do
            sum ← sum + c
        EMIT(edge e, sum)
```

Sử dụng 02 kỹ thuật là tổng hợp cục bộ (Combiner) trong lớp Mapper và duy trì trạng thái biến nhớ trên toàn bộ các tác vụ map, để tối ưu hóa lớp Mapper.

Bài làm:

```
Class MAPPER
Method INITIALIZE:
    H <- new ASSOCIATIVEARRAY()

Method MAP(polygonId i, polygon p)
    for all edge e ∈ polygon p do
        H{e} <- H{e} + 1

Method CLOSE
    for all edge e ∈ H do
        Emit (edge e, count H{1})
```

```
class REDUCER
    method REDUCE(edge e, counts[c1, c2,…])
        sum ← 0
        for all count c ∈ counts[c1, c2,…] do
            sum ← sum + c
        EMIT(edge e, sum)
```